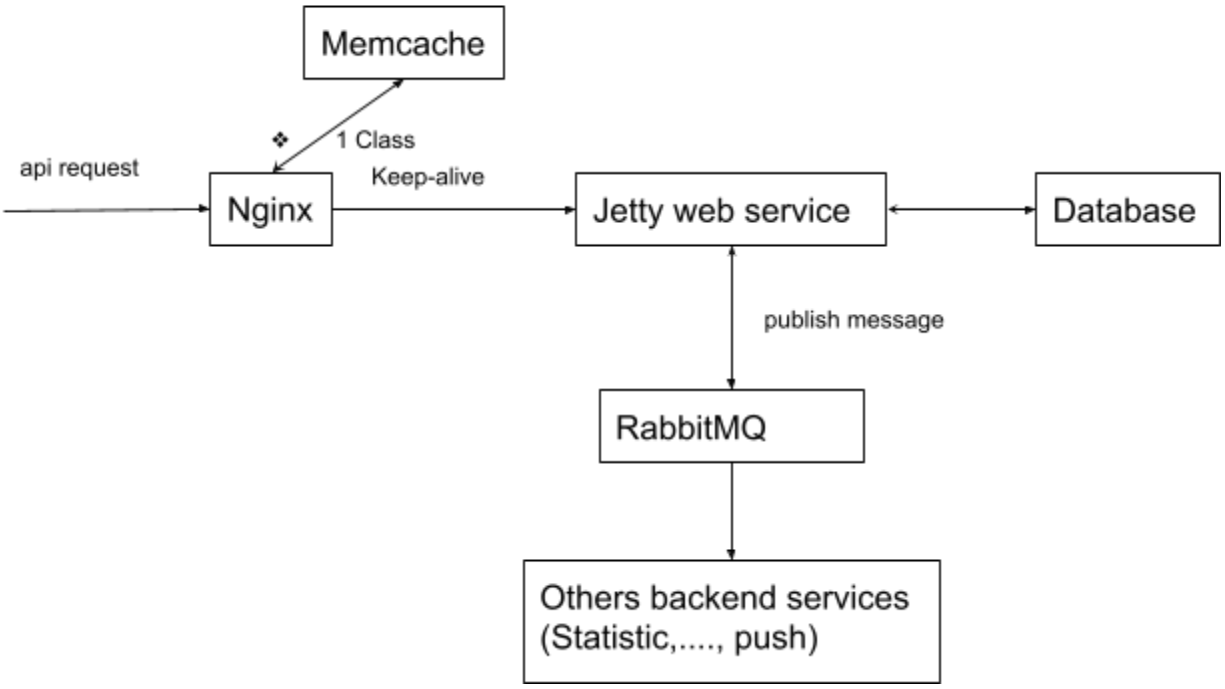
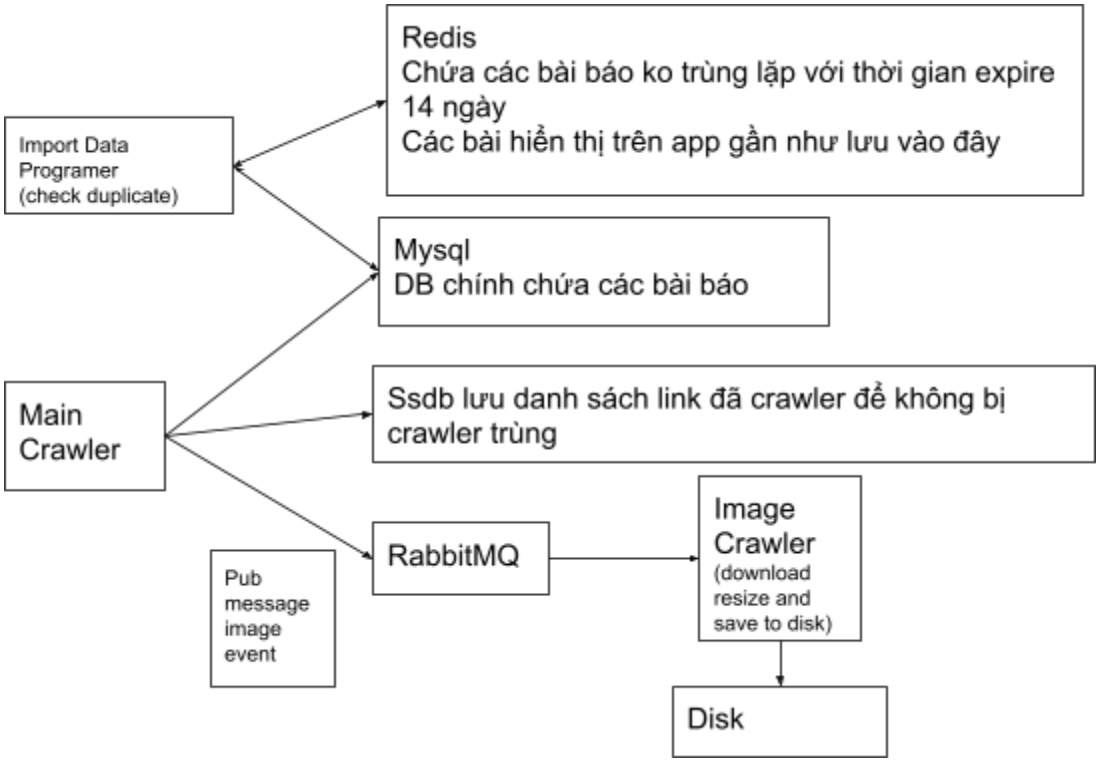
Kiến trúc hệ thống

1. Tổng quan





2. Miêu tả tổng quan

   Đọc báo 24h là ứng dụng tổng hợp tin tức từ ~70 đầu báo online hàng đầu. Hệ thống sẽ tự động lấy về những bài báo mới, các bài báo sẽ được hiển thị theo thời gian publish bài báo gốc.

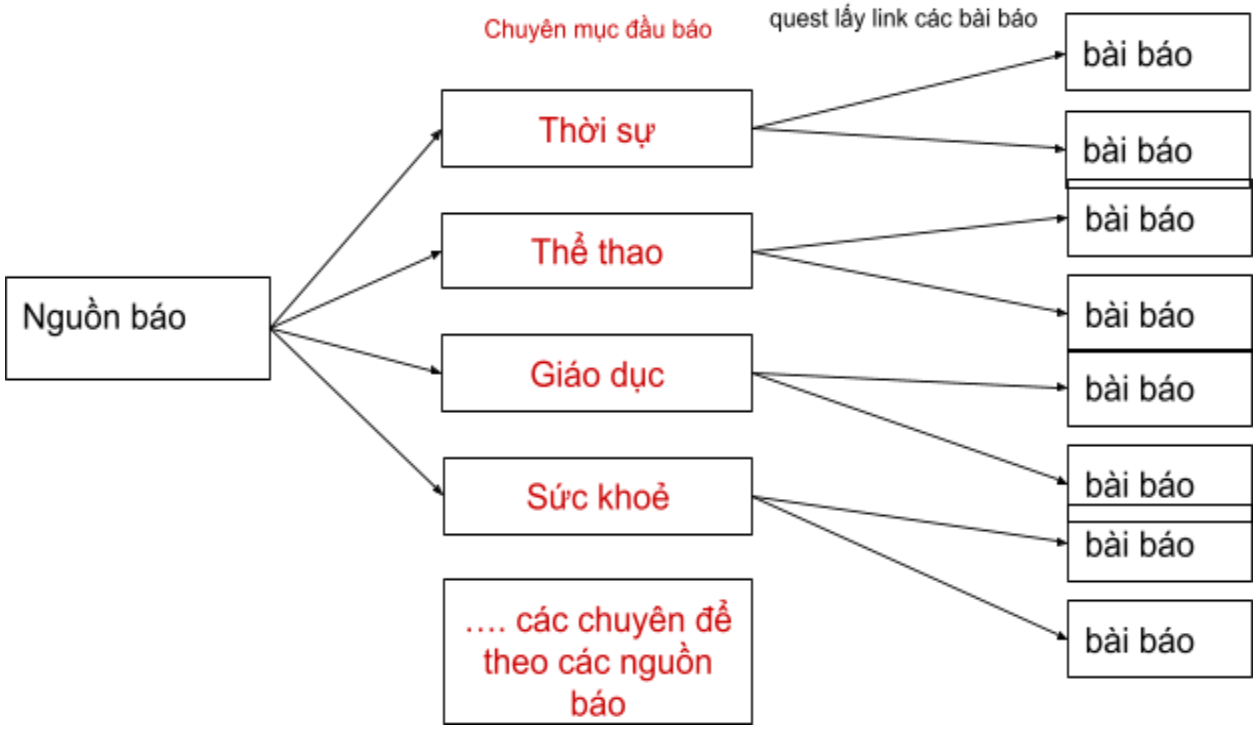
   Bản thân Đọc báo 24h sẽ được coi là 1 báo tổng hợp từ ~70 đầu báo online khác, bao gồm các chuyên mục :

   Mới nhất: tổng hợp tất cả các bài báo từ các nguồn báo
   Thời Sự
   Thể Thao
   Pháp luật
   Giải trí
   Tâm Sự
   Công nghệ
   Kinh tế
   Giáo dục
   Sức khoẻ
   Khám phá
   Xe cộ

Game
Cẩm nang
Cộng đồng

Các chuyên mục này có thể thay đổi dựa theo nhu cầu vận hành sản phẩm. Các bài báo từ các nguồn sẽ được chia vào các chuyên mục trên.

Các nguồn báo sẽ được tự động quét bài mới liên tục. Các link bài báo mới được lấy theo các chuyên mục chính



3. Các module chính
   Module crawler: Tự động định kì quét các chuyên mục của các nguồn báo, các link bài báo mới sẽ được lấy từ các chuyên mục của báo.

   Ví dụ với báo Dân Trí: Từ link chuyên mục xã hội

https://dantri.com.vn/xa-hoi.htm sẽ tìm ra các link bài báo -> từ danh sách link bài báo sẽ crawler thông tin chi tiết của mỗi bài, những link đã được crawler thành công sẽ được lưu trữ vào **SSDB** để tránh crawler lại bài báo ở những lần quét dữ liệu sau.
Các bài báo sẽ lấy các thông tin:
1. Title bài báo
2. Thumb bài báo
3. Sapo miêu tả về bài báo.
4. Nội dung bài báo (thẻ html chính chưa main content), ở bước crawler này sẽ chưa parse html.
5. Thời gian publish bài báo ---> thông tin này rất quan trọng vì sẽ là score của bài báo luôn, bài nào mới xuất bản sẽ được hiển thị ở phía trên.

Các link ảnh của bài báo bao gồm thumb + ảnh trong bài sẽ được gửi event (bằng rabbitMQ) cho 1 service download ảnh riêng.

Các nội dung chi tiết crawler bài báo sẽ được lưu trữ trong mysql.

Module Import dữ liệu: Tự động và định kì lấy những bài báo mới crawler về từ mysql, parser các thẻ html trong content bài báo, sau đó insert vào reddis. Thời gian expire trong redis là 2 tuần (thời gian này điều chỉnh phụ thuộc vào muốn bài báo tồn tại trong hệ thống bao lâu).
Mối bài báo sẽ được lưu bằng 1 key trong redis. Ngoài ra thông tin bài báo còn được thêm vào các sorted list theo các chuyên mục của nguồn báo.

Module DownloadImage: nhận event thông tin các ảnh cần download từ module crawler thông qua rabbitMQ. Tiến hành download và lưu xuống ổ cứng. Những file thumb thì sẽ resize kích thước.

Module VideoService: service để lấy lại link video trong bài với những trang họ để expire time video, phần video này là phiền phức nhất và các bên liên tục thay đổi các công nghệ về play video. Với những video nguồn từ youtube trong bài báo thì hệ thống đang ko xử lý.

Module api service: cung cấp phương thức để client lấy dữ liệu. Trực tiếp lấy dữ liệu từ redis. Các api gọi lên server đều có param “db24h” param này được generate dựa trên thuật toán của google authen. Mục tích là để các bên thứ 3 không thể lấy lại được nội dung báo của Đọc báo 24h. Các api quan trọng:

1. ***/docbao24h/api/v1.0/website*** : trả về thông tin config danh sách các đầu báo đang có trong hệ thống.

2. ***/docbao24h/api/v1.0/articles*** : tham số là websiteID + topicID → trả về danh sách các bài báo trong cùng chuyên mục của đầu báo.

3. ***/docbao24h/api/v1.0/articles/info*** : tham số là articleID → Trả về nội dung chi tiết của bài báo.

4. ***/docbao24h/api/v1.0/articles/relative*** : tham số là articleID → trả về danh sách bài báo liên quan của 1 bài báo .

5. ***/docbao24h/api/v1.0/articles/tintaitro***: trả về thông tin config những quảng cáo của bên thứ 3 đang triển khai trên app.